# EVREAL: Towards a Comprehensive Benchmark and Analysis Suite for Event-based Video Reconstruction

## Supplementary Material

Burak Ercan [1,2] Onur Eker [1,2] Aykut Erdem [3,4] Erkut Erdem [1,4]

[1] Hacettepe University, Computer Engineering Department [2] HAVELSAN Inc.
[3] Koç University, Computer Engineering Department [4] Koç University, KUIS AI Center

In this supplementary document, we provide additional material to complement the main paper. First, we present recent related work on event-based video reconstruction, especially focusing on their evaluation details (Sec. 1). Second, we share the details of the event representation that we have employed (Sec. 2). Third, we share implementation details of the evaluation metrics we considered in our analysis (Sec. 3). Next, we provide the overview of the datasets being used in our proposed EVREAL framework (Sec. 4). Then, we present the details and results from the computational complexity analysis that we performed (Sec. 5). Finally, we share additional qualitative results from several datasets (Sec. 6).

## 1. Related Work

In recent years, there has been a surge of methods aiming to reconstruct intensity images from events, each taking into account different assumptions and employing distinct processing methodologies. Early approaches were limited, often relying on basic assumptions such as known or restricted camera movement, static scenes, or brightness constancy. More recent methods utilize deep neural networks and incorporate natural image priors in their models to achieve better results. Here, we limit our discussion to these recent methods and especially focus on their evaluation details.

Wang *et al*. [33] proposed a conditional GAN based model, in which input events are represented by means of spatio-temporal voxel grids. Their evaluation setup includes a small amount of data containing 1000 intensity frames taken from both real and simulated datasets, including the sequences from [1]. They compared their method against [1, 17] for sequences without any ground truth outputs, by utilizing the no-reference metric BRISQUE [14]. The authors do not share their evaluation code.

Rebecq *et al*. [23, 24] introduced a recurrent fully convolutional network, named E2VID. The authors used a selection of seven sequences from the ECD [16] dataset, using a fixed number of events to form event voxel grids and a tolerance of 1 ms to match the reconstructions with ground truth frames. To improve the output quality, they applied robust normalization as a post-processing step and then performed local histogram equalization before computing scores for MSE, SSIM and LPIPS [37]. They compared their approach against [1] and [17]. They also reported a temporal consistency score that requires a ground truth optical flow map between each frame. To obtain this, they used an off-the-shelf frame-based optical flow network [11], which has its own prediction errors. The researchers conducted experiments on challenging scenarios involving rapid motion, low-light conditions and high dynamic range, without providing any quantitative scores. Additionally, they reported color image reconstruction results from the event data available in CED dataset [27], without providing any quantitative analysis.

Rebecq *et al*. also evaluated their method on four downstream tasks, including image classification, visual-inertial odometry, object detection, and monocular depth estimation [23, 24]. To perform these tasks, they fed reconstructed frames as inputs to task-specific frame-based methods and reported either qualitative or quantitative results. For instance, for object classification, they used events from N-MNIST [18], N-Caltech101 [18], and N-Cars [28] datasets, and provided accuracy scores achieved by a ResNet-18 [10] network. Similarly, for visual-inertial odometry, they employed events from the ECD dataset, and investigated mean translation errors obtained via VINS-Mono [21]. For object detection and monocular depth estimation, they used YOLOv3 [25] and MegaDepth [13], respectively, and only shared qualitative results in a supplementary video. Additionally, they analyzed the computational efficiency of their approach by reporting the frame synthesis time. The authors do not release their evaluation code publicly.

Scheerlinck *et al*. [26] proposed FireNet, a lightweight

recurrent network, as a replacement for E2VID, and demonstrated that it can attain similar performance with much less memory consumption and faster inference. In their evaluation setup, they followed the methodology in [24], and performed experiments on the selected frames from the sequences in the ECD dataset. They utilized a fixed number of events to form event voxel grids, and applied local histogram equalization to reconstructions and ground truth frames before estimating quantitative metrics such as MSE, SSIM, and LPIPS. Additionally, they performed qualitative analysis on color image reconstruction and challenging scenarios involving high-dynamic range and fast motion. They focused on evaluating computational efficiency and compared several resolutions on GPU and CPU by examining the number of model parameters, memory consumption, FLOPs, and inference times. However, they did not conduct any downstream task experiments, and their evaluation codes are not made publicly available.

Stoffregen *et al*. [30] proposed an enhanced version of E2VID, named E2VID+, by retraining it on synthetic training data exhibiting similar statistics with real-world test data. They also employed the same strategy for improving the FireNet architecture, resulting in FireNet+. They evaluated their methods on a larger set of real-world sequences from three datasets, namely ECD and MVSEC [39] datasets, and their proposed HQF dataset. For ECD and MVSEC, they used the sequences commonly used in earlier work, and reported MSE, SSIM, and LPIPS scores. They always had a matching ground truth frame for each reconstruction, as they used events between each consecutive ground truth frame to form voxel grids. It is not clear whether they applied normalization or histogram equalization before calculating these scores. Moreover, they did not perform any experiments on challenging scenarios or downstream tasks, nor did they perform computational efficiency analysis. The evaluation code is not publicly available.

Cadena *et al*. [4] proposed SPADE-E2VID, which integrates spatially-adaptive denormalization (SPADE) [20] layers into the E2VID architecture to enhance the quality of the reconstructed videos. The authors evaluated their approach using seven sequences from the ECD dataset, starting from the very first frames of each sequence, and reported MSE, SSIM, and LPIPS scores for quantitative comparison with E2VID and FireNet. They also introduced an RMS contrast metric to demonstrate that their method produces higher contrast reconstructions. To assess temporal consistency, they used a different off-the-shelf frame-based optical flow network [22] and reported the corresponding scores. In addition, they performed object detection analysis on a single sequence of the ECD dataset, using events and YOLOv4 [2] to process reconstructed frames. They estimated ground truth object labels for two object classes by applying the same object detection network to ground truth intensity images and shared average precision scores for this downstream task accordingly. They analyzed the computational efficiency of their approach by reporting reconstruction time for inputs with various resolutions. While they released an evaluation code, we were unable to reproduce their results with it.

Weng *et al*. [35] improved the E2VID architecture by adding a Transformer-based module to better exploit the global context of event tensors, thus naming their model as ET-Net. Their experiments were conducted using the ECD, MVSEC, and HQF datasets, with the same sequence cuts as in [30]. In their experiments, events between consecutive ground truth frames were used to form voxel grids. To evaluate their approach, they calculated MSE, SSIM, and LPIPS scores, without any normalization or histogram equalization applied to the reconstructed images. They compared their method with E2VID, E2VID+, FireNet, and FireNet+, and shared qualitative results on challenging scenarios involving high-dynamic-range and rapid motion in their supplementary material. However, they did not perform a computational efficiency analysis or an experiment on a downstream task. The authors provided an open-source evaluation code, and we are able to use to reproduce their results.

Paredes-Vallés and de Croon [19] proposed a self-supervised learning method called SSL-E2VID, which employs the event-based photometric constancy assumption [8] to estimate optical flow and intensity images simultaneously. As done in earlier work, events between each consecutive ground truth frame were used to form voxel grids. Their experiments were conducted on ECD and HQF datasets, and they made quantitative comparisons with E2VID, E2VID+, FireNet, and FireNet+. Local histogram equalization was employed before calculating quantitative scores. Since they did not introduce a new architecture, computational efficiency analysis was not performed. Qualitative results were given also for challenging scenarios such as high-dynamic-range and high-speed. No downstream task analysis was performed, and their evaluation code was not made publicly available.

Zhu et al. [40] proposed a spiking neural network architecture that achieves comparable performance to E2VID, E2VID+, FireNet, and SPADE-E2VID with higher computational efficiency. They used the ECD, MVSEC, and HQF datasets in their evaluation and reported quantitative scores using MSE, SSIM, and LPIPS metrics. In reconstructing intensity images, they used the events between each consecutive ground truth frame as input. They applied histogram equalization before calculating these scores. In addition, they provided an analysis of energy consumption. However, they did not release an open-source evaluation code.

Zhang *et al*. [38] presented a novel approach for event-based image reconstruction by formulating it as a linear inverse problem based on optical flow. They conducted a

quantitative comparison with E2VID, E2VID+, and SSL-E2VID using MSE, SSIM, and LPIPS metrics. They focused on test sequences with limited camera motion, specifically selected from the ECD dataset, and utilized events from N-Caltech101 [18] dataset. They aligned reconstructions with respective reference frames using Enhanced Correlation Coefficient Maximization [6]. They reported median scores for each sequence instead of mean scores and presented distribution plots of scores of each method on various sequences. They also analyzed the effect of histogram equalization on quantitative scores and emphasized the importance of taking various factors into account while interpreting these scores. They showcased their method's ability to perform color reconstruction and demonstrated temporal consistency on two example frames from the DSEC dataset [9]. They did not conduct experiments on downstream tasks and did not share their evaluation code.

## 2. Details of Event Representation

Following the common practice in the literature, we form voxel grids from grouped events in order to utilize deep CNN architectures for event-based data. Let $G_k$ be a group of events that span a duration of $\Delta T$ seconds, $T_k$ be the starting timestamp of that duration, and $B$ be the number of temporal bins used to discretize the timestamps of continuous-time events in the group. The voxel grid $V_k \in \mathbb{R}^{W \times H \times B}$ for that group is formed by normalizing the timestamps of events from the group to the range $[0, B-1]$. Each event contributes its polarity to the two temporally closest voxels using a linearly weighted accumulation similar to bilinear interpolation. Specifically, the voxel grid is computed as follows:

$$V_k(x, y, t) = \sum_i p_i \max(0, 1 - |t - t_i^*|)\delta(x - x_i, y - y_i) \tag{1}$$

where $\delta$ is the Kronecker delta that selects the pixel location, and $t_i^*$ is the normalized timestamp which is calculated as:

$$t_i^* = (B-1)(t_i - T_k)/(\Delta T) \tag{2}$$

In all our experiments, we set the number of temporal bins $B$ to 5.

## 3. Implementation Details for Quantitative Image Quality Metrics

**MSE.** The Mean Squared Error is a commonly used metric that does not require any parameters. When comparing two images, the only factor that can impact the MSE result is the range of pixel values that the images possess. We calculate the MSE using floating-point pixel values within the range $[0, 1]$. Lower MSE values indicate better results.

**SSIM.** We utilize the `scikit-image` image processing library's [32] implementation for structural similarity. We adjust the parameters to use the Gaussian weighting scheme explained in [34]. Like MSE, we compute SSIM using images with floating point pixel values in the range of $[0, 1]$. Higher scores of SSIM indicate better results.

**LPIPS.** We utilize the official implementation of LPIPS [37][1], v0.1.4, and employ the variant that uses the pre-trained AlexNet [12] network. To comply with the implementation, we normalize the images so that their pixel values fall in the range of $[-1, 1]$. In the LPIPS score calculation, a lower score indicates better quality.

**BRISQUE.** For BRISQUE [14], we use the implementation in IQA-PyTorch toolbox [5][2], v0.1.5, with default settings. The implementation supports 3-channel RGB images. Therefore, we convert intensity images into RGB images by concatenating three copies of the grayscale image along the third channel before calculating the scores.

**NIQE.** For NIQE [15], we again use the implementation in IQA-PyTorch toolbox [5], v0.1.5, with default settings. The implementation supports 3-channel RGB images. Therefore, we convert intensity images into RGB images by concatenating three copies of the grayscale image along the third dimension before calculating the scores.

**MANIQA.** For MANIQA [36], we also use the implementation in IQA-PyTorch toolbox [5], v0.1.5, with default settings. The implementation supports 3-channel RGB images. Therefore, we convert intensity images into RGB images by concatenating three copies of the grayscale image along the third channel before calculating the scores. MANIQA works by taking random crops of size $224 \times 224$ pixels from the images, whereas the ECD dataset used in our analysis has a lower resolution. To address this discrepancy, we upscale the images to the desired size before calculating the scores.

## 4. Dataset Details

**Event Camera Dataset (ECD).** Following the practice explained in [24], we use seven different sequence with diverse characteristics from the ECD dataset [16]. These sequences are short, taken indoors, and mostly contain simple scenes of office environments with stable objects. The data was captured by a DAVIS240C sensor [3], which is mostly moving with 6 degrees of freedom (DOF) and with increasing speed. The camera generates events and frames from the same pixel array, which has a spatial resolution of $240 \times 180$

[1]The code is accessible from https://github.com/richzhang/PerceptualSimilarity

[2]The code is accessible from https://github.com/chaofengc/IQA-PyTorch

pixels. The ground truth intensity frames are available at an average rate of 22 Hz.

To allow methods to generate meaningful results, we exclude the initial few seconds of each sequence from quantitative evaluation. Additionally, when using full-reference metrics, as commonly done in earlier work, we do not include the latter parts of the sequences as they may contain motion blur due to the increased speed of camera movement. However, when evaluating with no-reference metrics, we specifically concentrate on these sections that have fast camera movement, to which the corresponding ground truth intensity images are of lower quality.

**Multi Vehicle Stereo Event Camera (MVSEC) dataset.** The MVSEC dataset [39] contains longer sequences captured by a pair of DAVIS 346B cameras, each having a spatial resolution of $346 \times 260$ pixels. These sequences depict both indoor and outdoor environments. To evaluate the quality of the videos generated by the methods using full-reference metrics, we followed the approach taken by [30] and considered six commonly used sequences from this dataset. Four of these sequences were captured indoors by a flying hexacopter, while the remaining two were taken outdoors during the daytime from a driving vehicle. The average rate of ground truth intensity frames was approximately 30 Hz for indoor sequences and 45 Hz for outdoor sequences. Additionally, we used three night sequences from this dataset, each captured from a vehicle as well, for our experimental evaluation involving no-reference metrics as the ground truth frames at night-time tend to be underexposed.

**High-Quality Frames (HQF) dataset.** The HQF dataset [30] contains fourteen sequences that exhibit a wide range of different motion behaviors, including static, slow, and fast camera motion, and cover both indoor and outdoor scenes. Two different DAVIS240C cameras are used to capture the data, providing distinct noise and contrast threshold characteristics. The cameras generate events and intensity frames from the same $240 \times 180$ pixel array. The scenes and camera parameters are adjusted to ensure that the ground truth frames are well-exposed and have minimal motion-blur. The average rate of ground truth intensity frames is 22.5 Hz. We use the entire sequences from this dataset for evaluation using full-reference quantitative metrics.

**Beam Splitter Event and RGB (BS-ERGB) Dataset.** The BS-ERGB Dataset [31] is originally collected for the event-based video frame interpolation task. The dataset consists of events recorded by a Prophesee Gen4M event camera [7] having a spatial resolution of $1280 \times 720$ pixels, and RGB frames captured by a global shutter RGB Flir camera with a resolution of $4096 \times 2196$ pixels. Both of these data are then post-processed to have the same spatial resolution of $970 \times 625$ pixels. Most of the sequences are short and captured with a static camera observing fast motions in the scene. Since events are confined to small regions where motion is observed, reconstructing intensity frames for other parts of the scene is not feasible. There are a few sequences recorded with a handheld camera where every pixel generates many events. We evaluate the models on ten of these handheld sequences.

| Network Architecture | Number of Params (M) | Inference Time (ms) |
|---|---|---|
| E2VID [19, 24, 30] | 10.71 | 5.1 |
| FireNet [26, 30] | 0.04 | 1.6 |
| SPADE-E2VID [4] | 11.46 | 16.1 |
| ET-Net [35] | 22.18 | 32.1 |

Table 1. **Computational complexity of different network architectures** in terms of the number of model parameters (in millions) and inference time (in milliseconds).

**High Speed and HDR Datasets** These high-speed and HDR sequences are recorded by Rebecq *et al*. [24], using a Samsung DVS Gen3 event camera [29] with a spatial resolution of $640 \times 480$. We use all three HDR sequences from this dataset, namely the $hdr\_selfie$, $hdr\_sun$, and $hdr\_tunnel$ sequences.

## 5. Computational Complexity

We also analyzed the computational complexity of each method by considering two metrics: the number of model parameters and inference time. The former is an essential metric as it indicates the memory requirements, while the latter reflects the real-time performance by determining the maximum FPS that can be achieved. To measure the inference time, we used a workstation equipped with a Quadro RTX 5000 GPU and considered data with a spatial resolution of $240 \times 180$. We report the average inference time for each method in ms. Table 1 compares the computational complexity of image reconstruction methods. In this table, we use the same row for the methods that share the same deep architecture. Overall, in terms of the number of parameters and inference times, FireNet is much smaller and faster than E2VID, while SPADE-E2VID is slightly larger and slower. ET-Net has the highest number of parameters which is twice as large as SPADE-E2VID, the second largest model, and its inference time is approximately $6\times$ slower than E2VID and $20\times$ slower than FireNet.

## 6. Additional Qualitative Results

Here, we provide qualitative comparisons for various sequences from the ECD, MVSEC, HQF, BS-ERGB, ECD-FAST, and MVSEC-NIGHT datasets. We present these results in Figures 1-6.

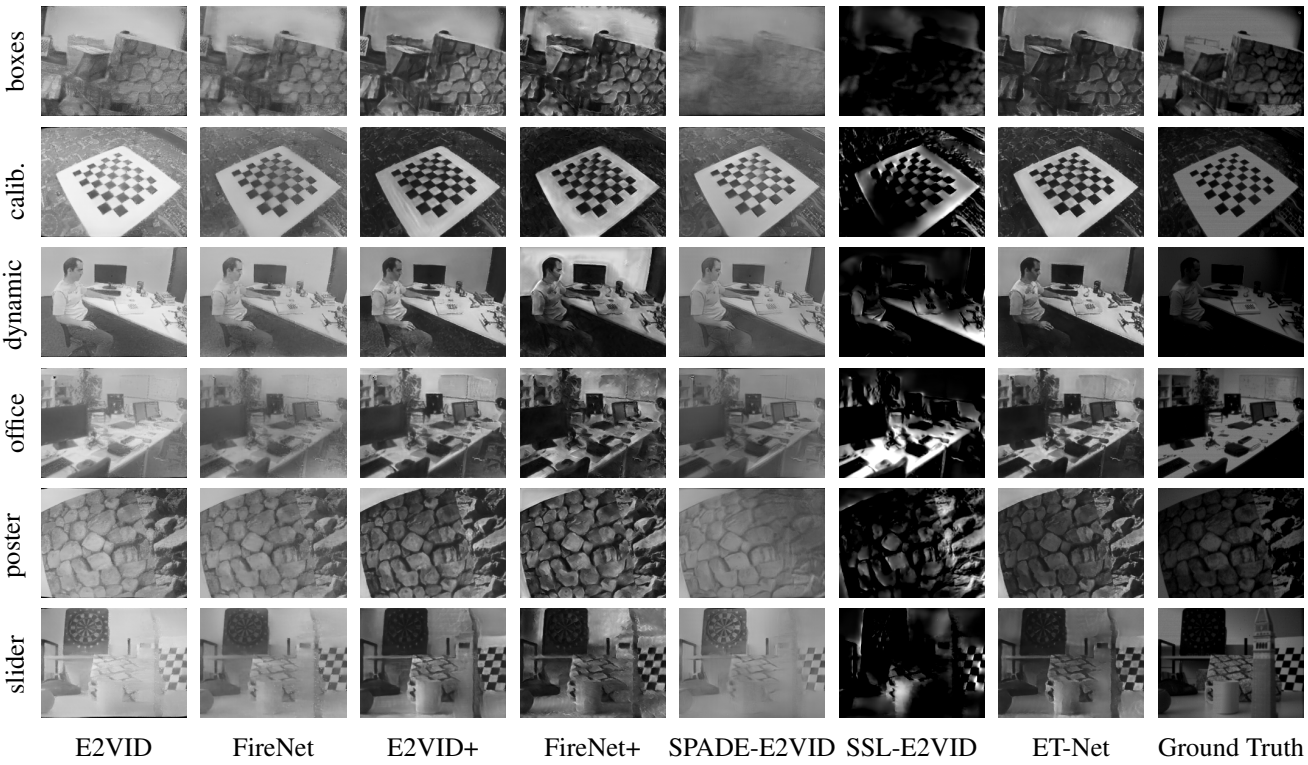


Figure 1. Additional qualitative comparisons on the ECD dataset.

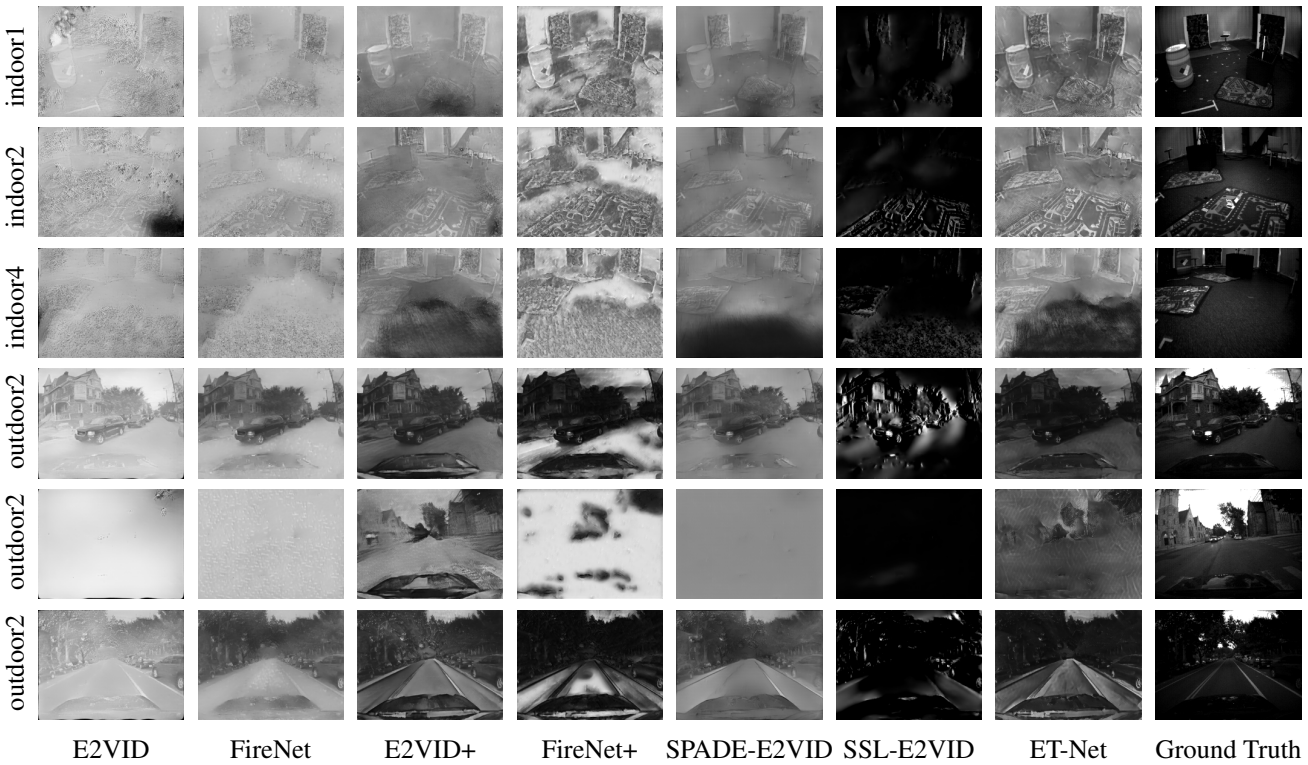


Figure 2. Additional qualitative comparisons on the MVSEC dataset.

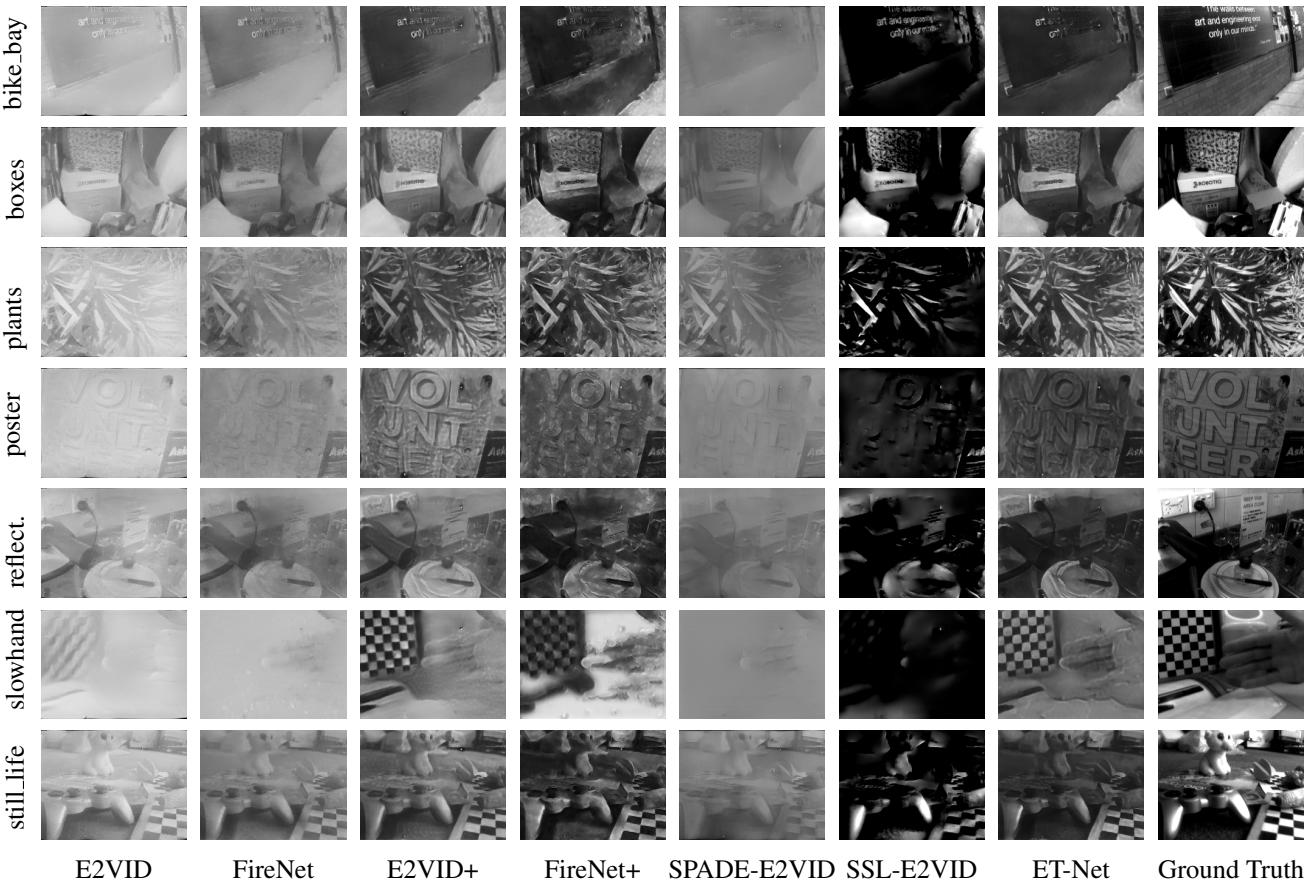


Figure 3. Additional qualitative comparisons on the HQF dataset.

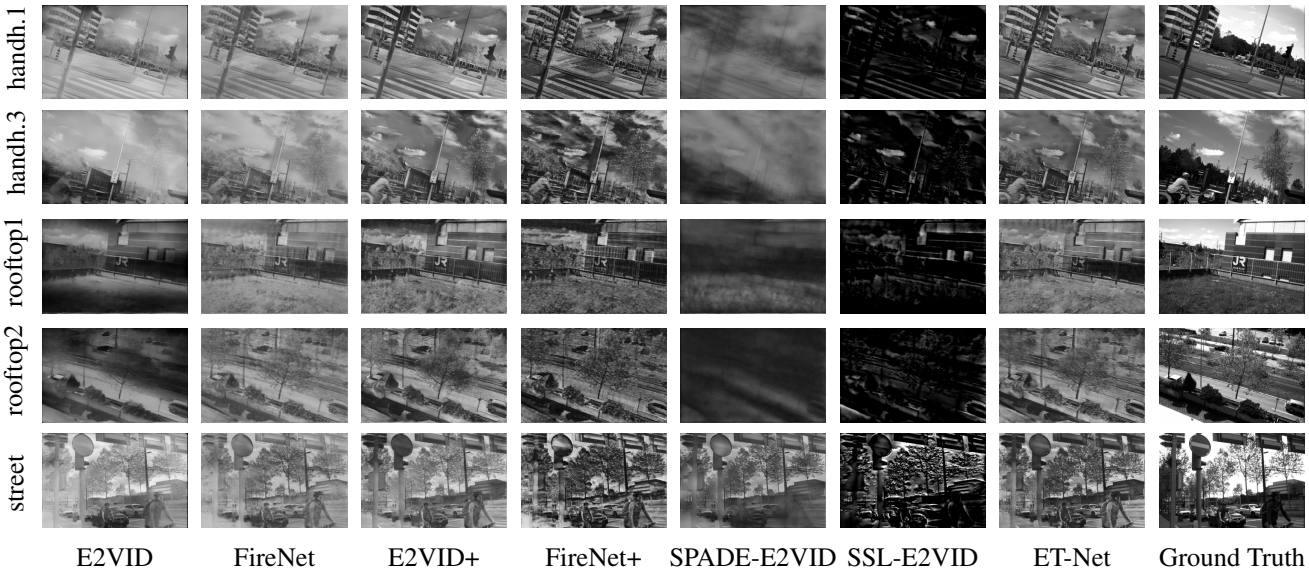


Figure 4. Additional qualitative comparisons on the BS-ERGB dataset.

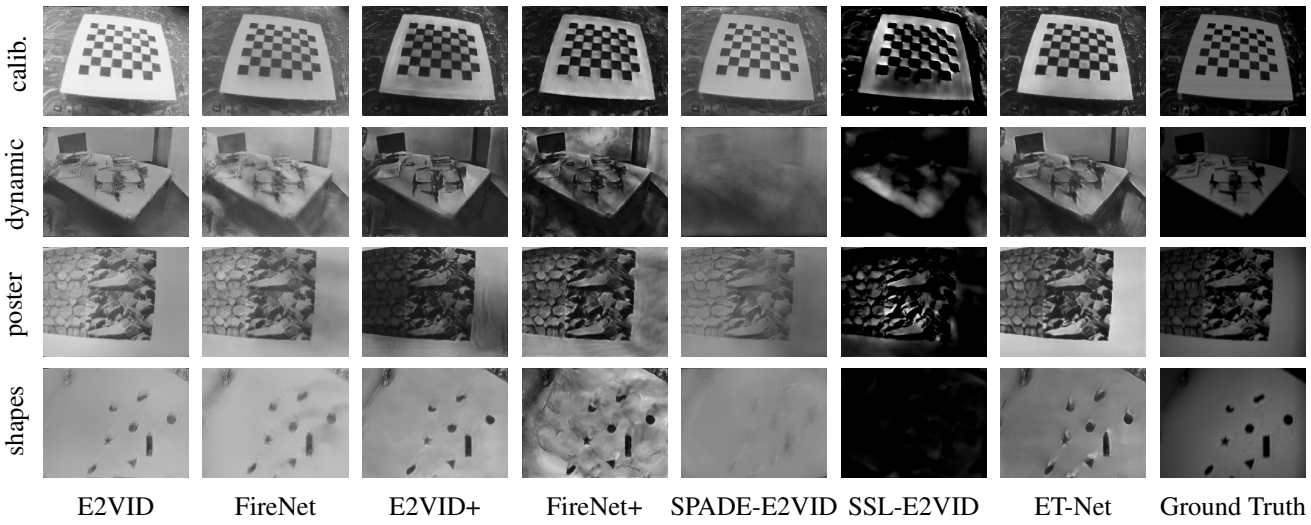


Figure 5. Additional qualitative comparisons on the fast parts of the ECD dataset (ECD-FAST).

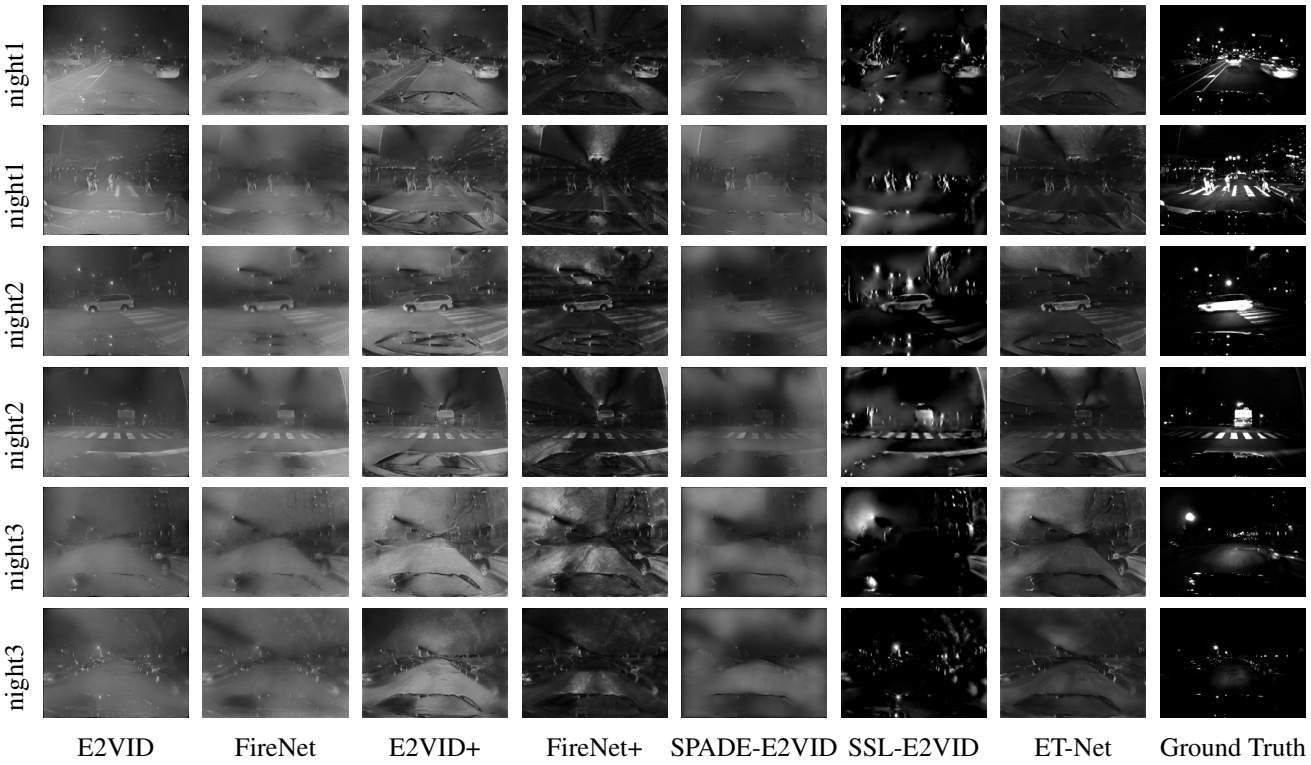


Figure 6. Additional qualitative comparisons on the night sequences of the MVSEC dataset (MVSEC-NIGHT).

## References


[1] Patrick Bardow, Andrew J Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 884–892, 2016. 1

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 2

[3] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240×180 130 db 3 $\mu$s latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 3

[4] Pablo Rodrigo Gantier Cadena, Yeqiang Qian, Chunxiang Wang, and Ming Yang. SPADE-E2VID: Spatially-adaptive denormalization for event-based video reconstruction. *IEEE Transactions on Image Processing*, 30:2488–2500, 2021. 2, 4

[5] Chaofeng Chen and Jiadi Mo. IQA-PyTorch: Pytorch toolbox for image quality assessment. [Online]. Available: https://github.com/chaofengc/IQA-PyTorch, 2022. 3

[6] Georgios D Evangelidis and Emmanouil Z Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE transactions on pattern analysis and machine intelligence*, 30(10):1858–1865, 2008. 3

[7] Thomas Finateu, Atsumi Niwa, Daniel Matolin, Koya Tsuchimoto, Andrea Mascheroni, Etienne Reynaud, Pooria Mostafalu, Frederick Brady, Ludovic Chotard, Florian LeGoff, et al. 5.10 a 1280×720 back-illuminated stacked temporal contrast event-based vision sensor with 4.86 $\mu$m pixels, 1.066 geps readout, programmable event-rate controller and compressive data-formatting pipeline. In *2020 IEEE International Solid-State Circuits Conference-(ISSCC)*, pages 112–114. IEEE, 2020. 4

[8] Guillermo Gallego, Christian Forster, Elias Mueggler, and Davide Scaramuzza. Event-based camera pose tracking using a generative event model. *arXiv preprint arXiv:1510.01972*, 2015. 2

[9] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[11] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 1

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 3

[13] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 1

[14] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 1, 3

[15] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 3

[16] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International Journal of Robotics Research*, 36(2):142–149, 2017. 1, 3

[17] Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *International Journal of Computer Vision*, 126(12):1381–1393, 2018. 1

[18] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015. 1, 3

[19] Federico Paredes-Vallés and Guido CHE de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3446–3455, 2021. 2, 4

[20] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 2

[21] Tong Qin, Peiliang Li, and Shaojie Shen. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. 1

[22] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017. 2

[23] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3857–3866, 2019. 1

[24] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1, 2, 3, 4

[25] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[26] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, pages 156–163, 2020. 1, 4

[27] Cedric Scheerlinck, Henri Rebecq, Timo Stoffregen, Nick Barnes, Robert Mahony, and Davide Scaramuzza. CED:

color event camera dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[28] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. HATS: histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1731–1740, 2018. 1

[29] Bongki Son, Yunjae Suh, Sungho Kim, Heejae Jung, Jun-Seok Kim, Changwoo Shin, Keunju Park, Kyoobin Lee, Jinman Park, Jooyeon Woo, et al. 4.1 a 640×480 dynamic vision sensor with a 9$\mu$m pixel and 300meps address-event representation. In *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 66–67. IEEE, 2017. 4

[30] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *European Conf. Comput. Vis.(ECCV)*, 2020. 2, 4

[31] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time Lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17755–17764, 2022. 4

[32] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in Python. *PeerJ*, 2:e453, 2014. 3

[33] Lin Wang, S. Mohammad Mostafavi, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10081–10090, 2019. 1

[34] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 3

[35] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2563–2572, 2021. 2, 4

[36] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 3

[37] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1, 3

[38] Zelin Zhang, Anthony Yezzi, and Guillermo Gallego. Formulating event-based image reconstruction as a linear inverse problem with deep regularization using optical flow. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, early access, Dec. 20, 2022. 2

[39] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3D perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018. 2, 4

[40] Lin Zhu, Xiao Wang, Yi Chang, Jianing Li, Tiejun Huang, and Yonghong Tian. Event-based video reconstruction via potential-assisted spiking neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3594–3604, 2022. 2